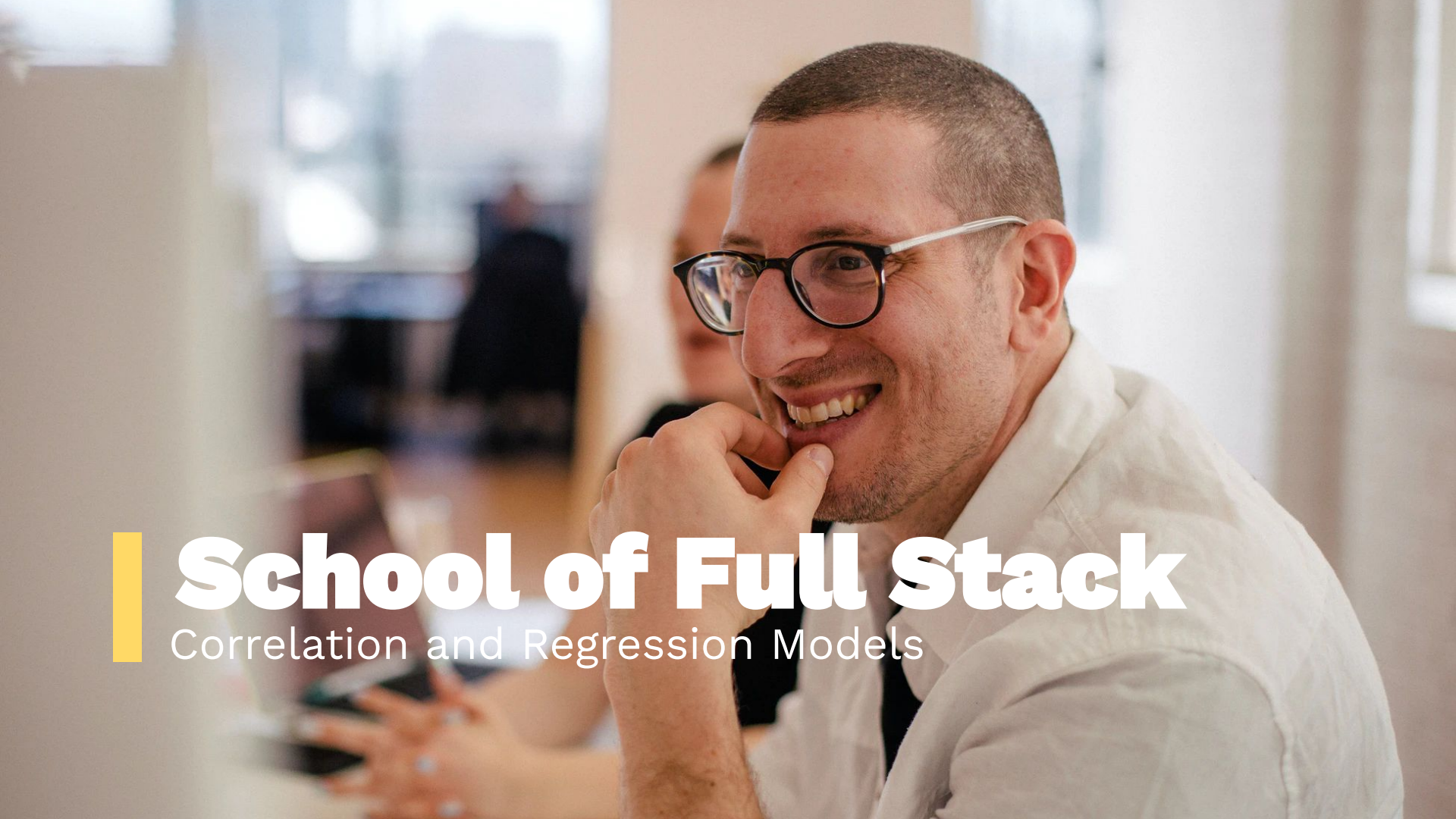
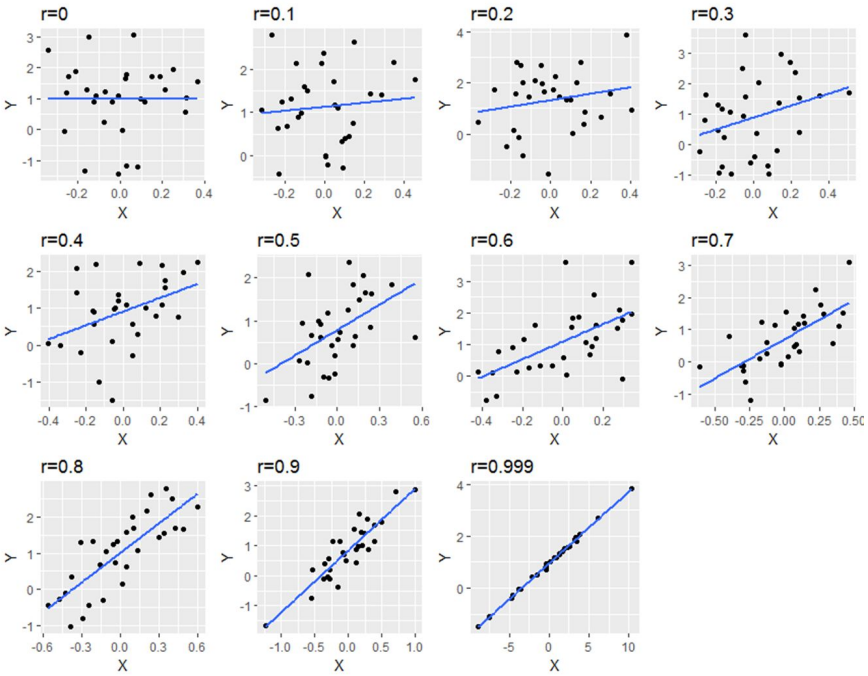


School of Full Stack

Correlation and Regression Models

Contents

What is Correlation

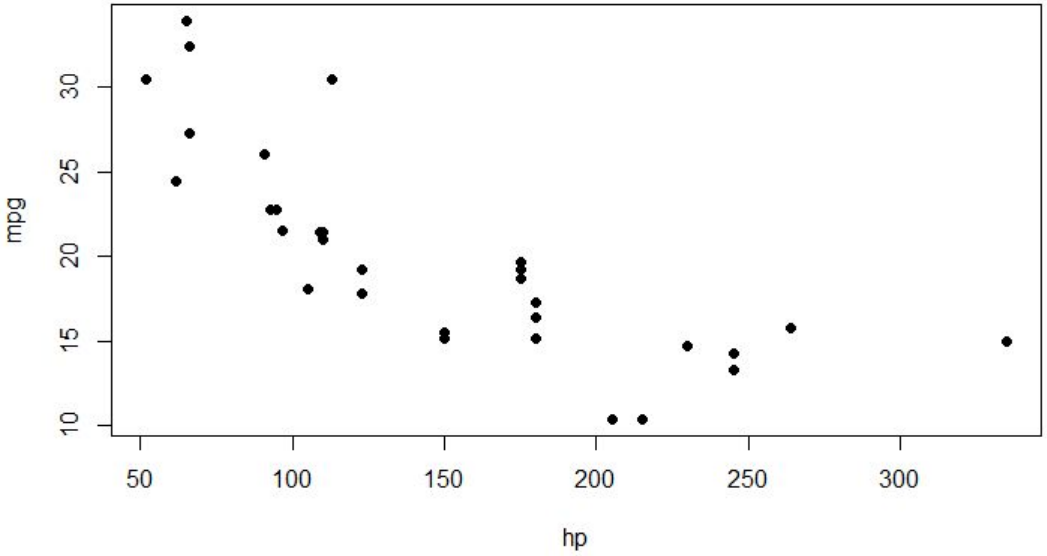


Correlation คือสถิติสำหรับหาความสัมพันธ์ของตัวแปร numeric สองตัว

- ค่าวิ่งอยู่ระหว่าง -1 , +1
- เครื่องหมาย -/+ บอกทิศทางความสัมพันธ์ของตัวแปรสองตัว
- ยิ่งค่าเข้าใกล้ | 1 | ความสัมพันธ์ยิ่งสูง

Explain Correlation

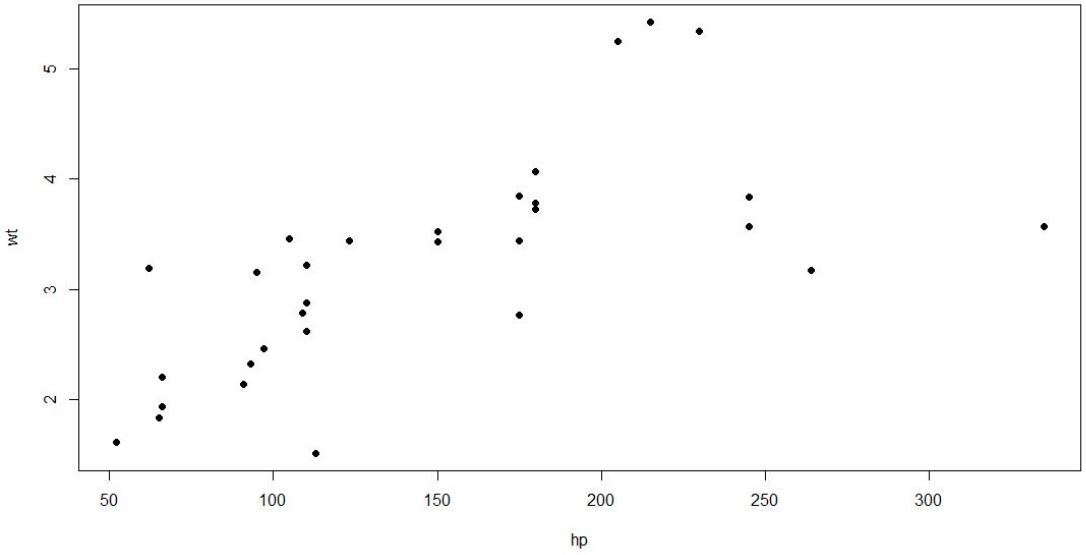
mile per gallon



horse power

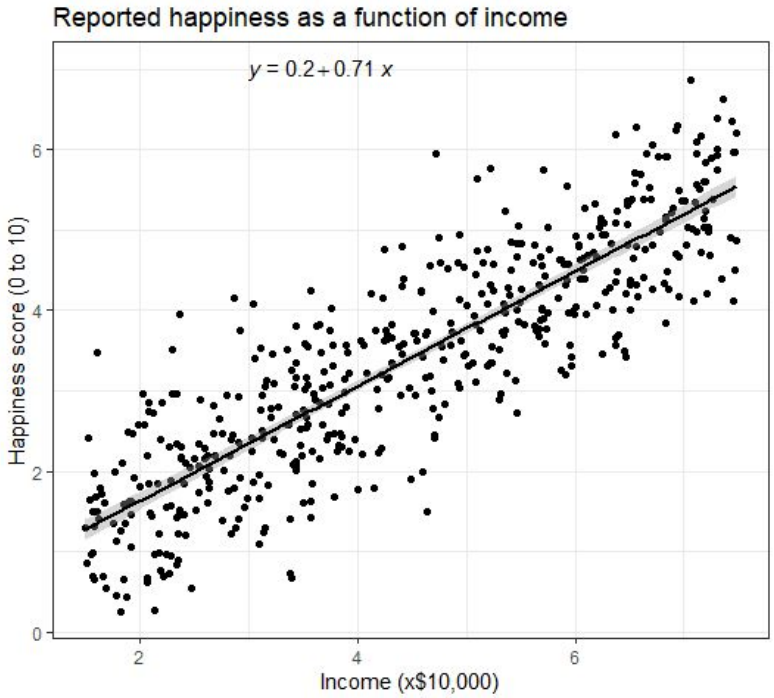
Explain Correlation

weight



horse power

What is Linear Regression



y = **intercept** + **slope***x

y = **b0** + **b1***x

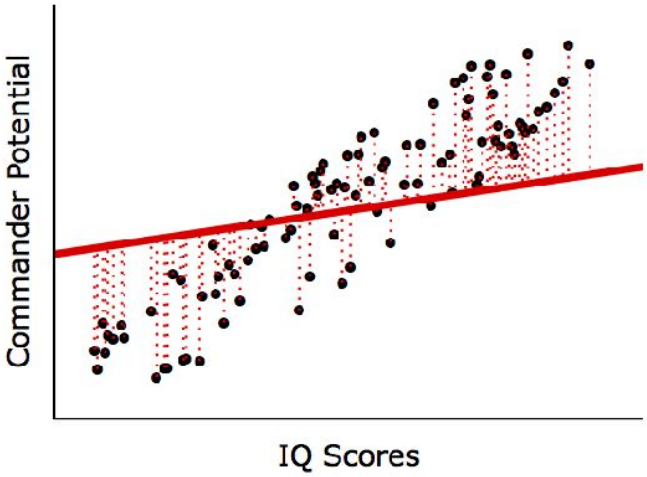
What is Linear Regression

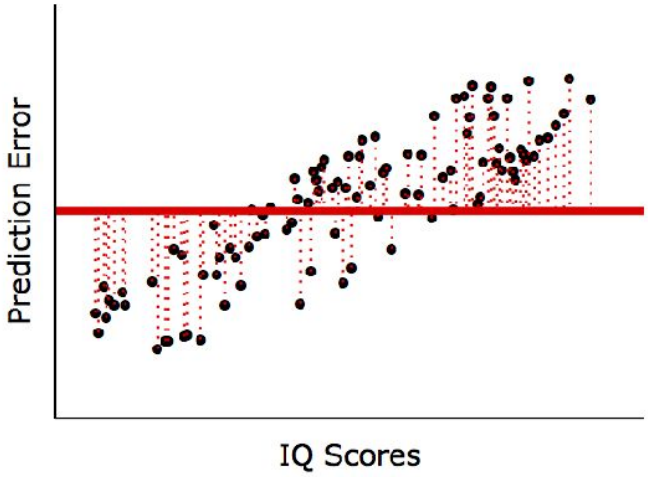
y = **intercept** + **slope***x

y = **b0** + **b1***x

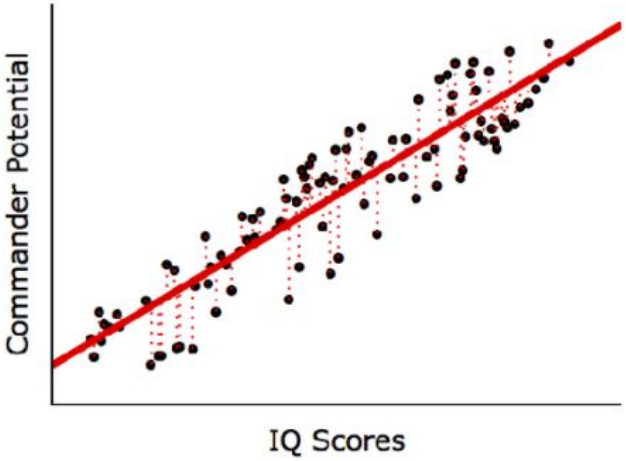
slope = Δy / Δx

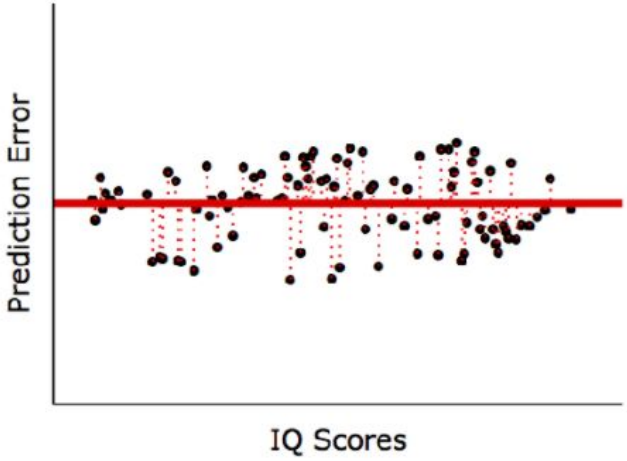
How the algorithm work





The best fitted line = lowest error





Knowledge Check

$$MPG = 30.09 - 0.06HP$$

1. Correlation ระหว่างตัวแปรสองตัวนี้เป็นแบบ positive หรือ negative
2. ถ้า HP=0 ค่า MPG จะเป็นเท่าไหร่
3. ถ้า HP=200 ค่า MPG จะเป็นเท่าไหร่

Basic Forms of Linear Regression

```
// Simple linear regression
y = b0 + b1*x1


// Multiple linear regression
y = b0 + b1*x1 + b2*x2 + b3*x3 + ... + bk*xk
```

Advanced - Normal Equation

Normal equation

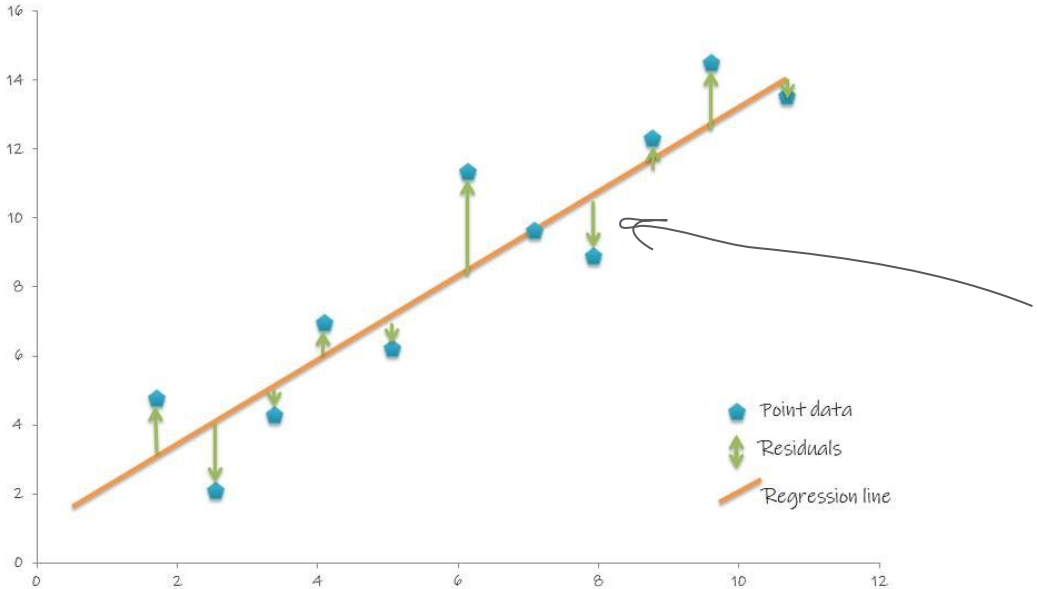
$$\Theta = (X^T X)^{-1} X^T y$$

Model Evaluation

Root Mean Squared Error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

RMSE lower is better

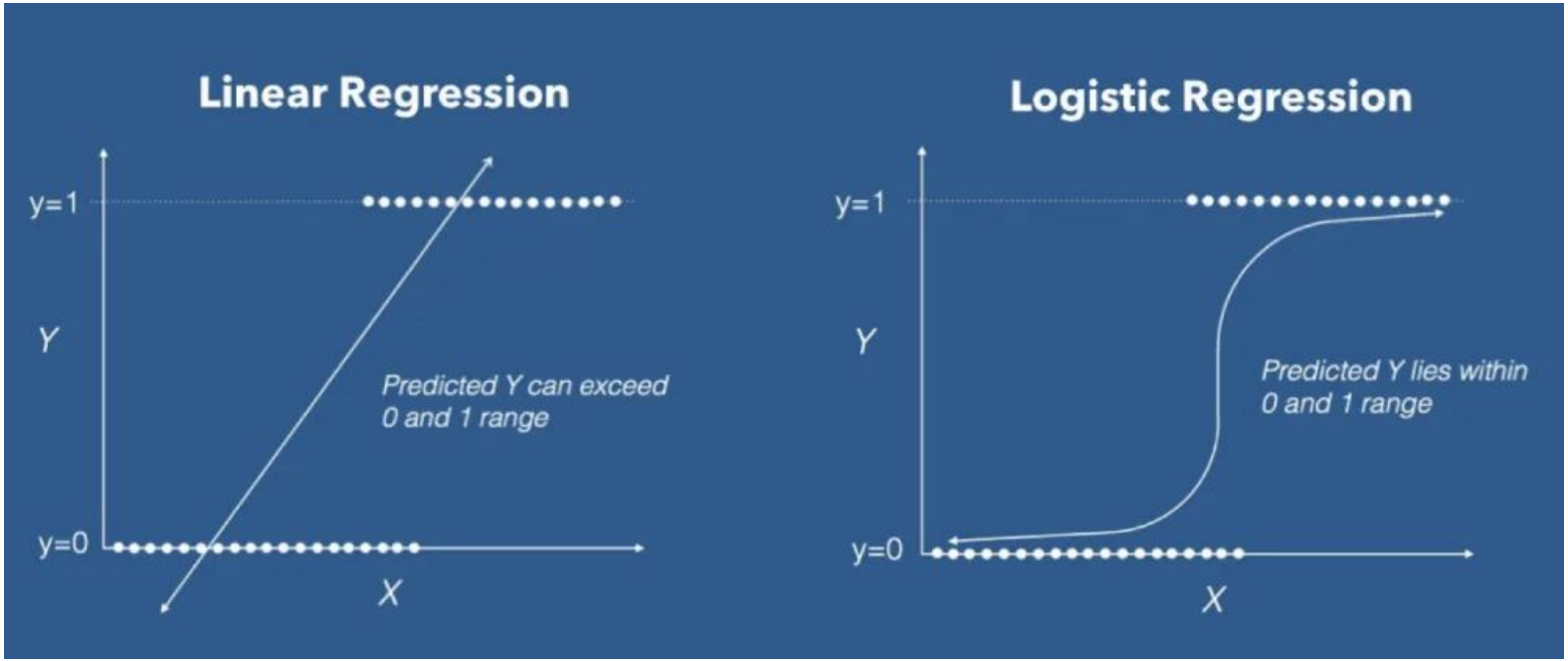


Error หรืออีกชื่อคือ Residual

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

[How to calculate the Root Mean Square Error (RMSE) of an interpolated pH raster? — Hatari Labs]

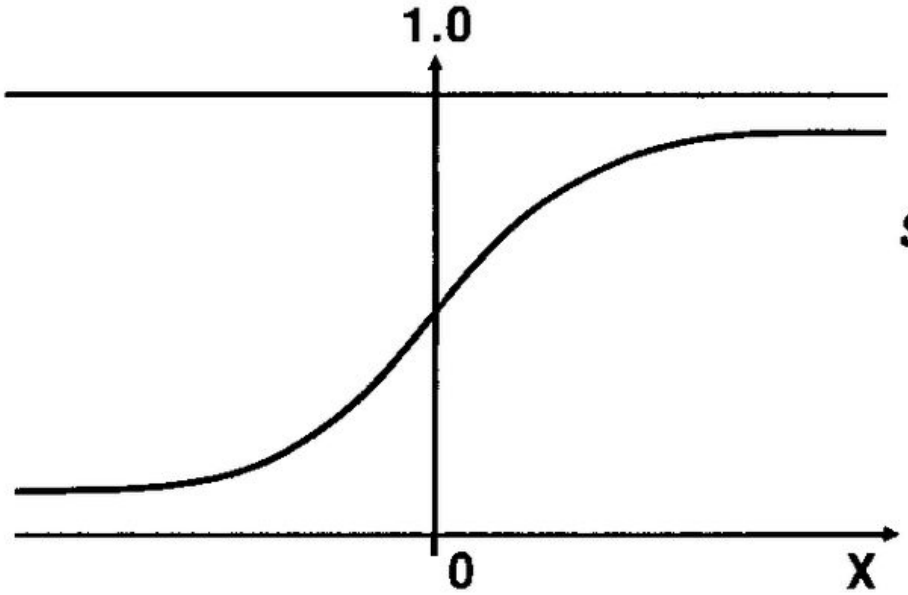
What is Logistic Regression



[Logistic Regression!!!! - DEV]

Logistic Function

เราใช้สูตรคณิตศาสตร์ เพื่อปรับผลลัพธ์ให้อยู่ระหว่าง 0-1 (เหมือนความน่าจะเป็น)



1

2

Sigmoid

$$f(x) = \frac{1}{1+e^{-x}}$$

e^z / (1 + e^z)

สองสูตรได้ผลลัพธ์เหมือนกัน

Model Evaluation - Confusion Matrix

		Actual	
		Yes	No
Predicted	Yes	50	16
	No	9	25

Model Evaluation - Confusion Matrix

		Actual	
		Yes	No
Predicted	Yes	50	16
	No	9	25

Accuracy

(50+25) / 100 = 75%

Model Evaluation - Confusion Matrix

		Actual	
		Yes	No
Predicted	Yes	50	16
	No	9	25

Recall

50 / (50+9) = 84.7%

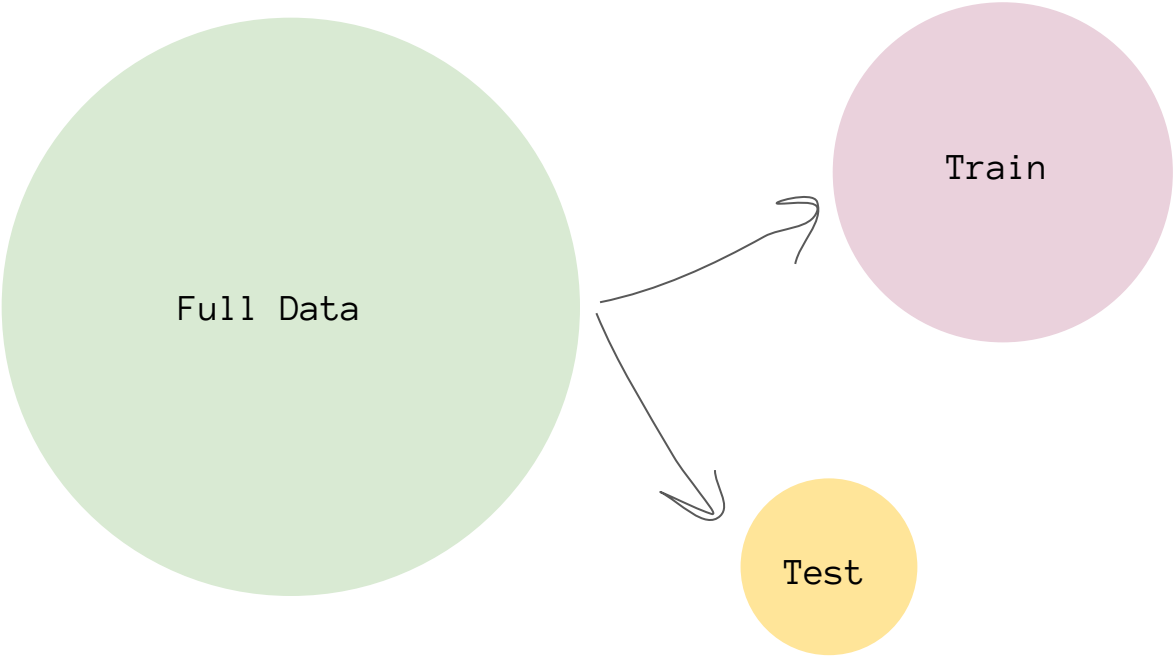
Model Evaluation - Confusion Matrix

		Actual	
		Yes	No
Predicted	Yes	50	16
	No	9	25

Precision

50 / (50+16) = 75.7%

Model Training Golden Rule



Correlation Plot

[An Introduction to corrplot Package (r-project.org)]

An Introduction to corrplot Package

Introduction

The **corrplot** package is a graphical display of a correlation matrix, confidence interval. It also contains some algorithms to do matrix reordering. In addition, corrplot is good at details, including choosing color, text labels, color labels, layout, etc.

Visualization methods

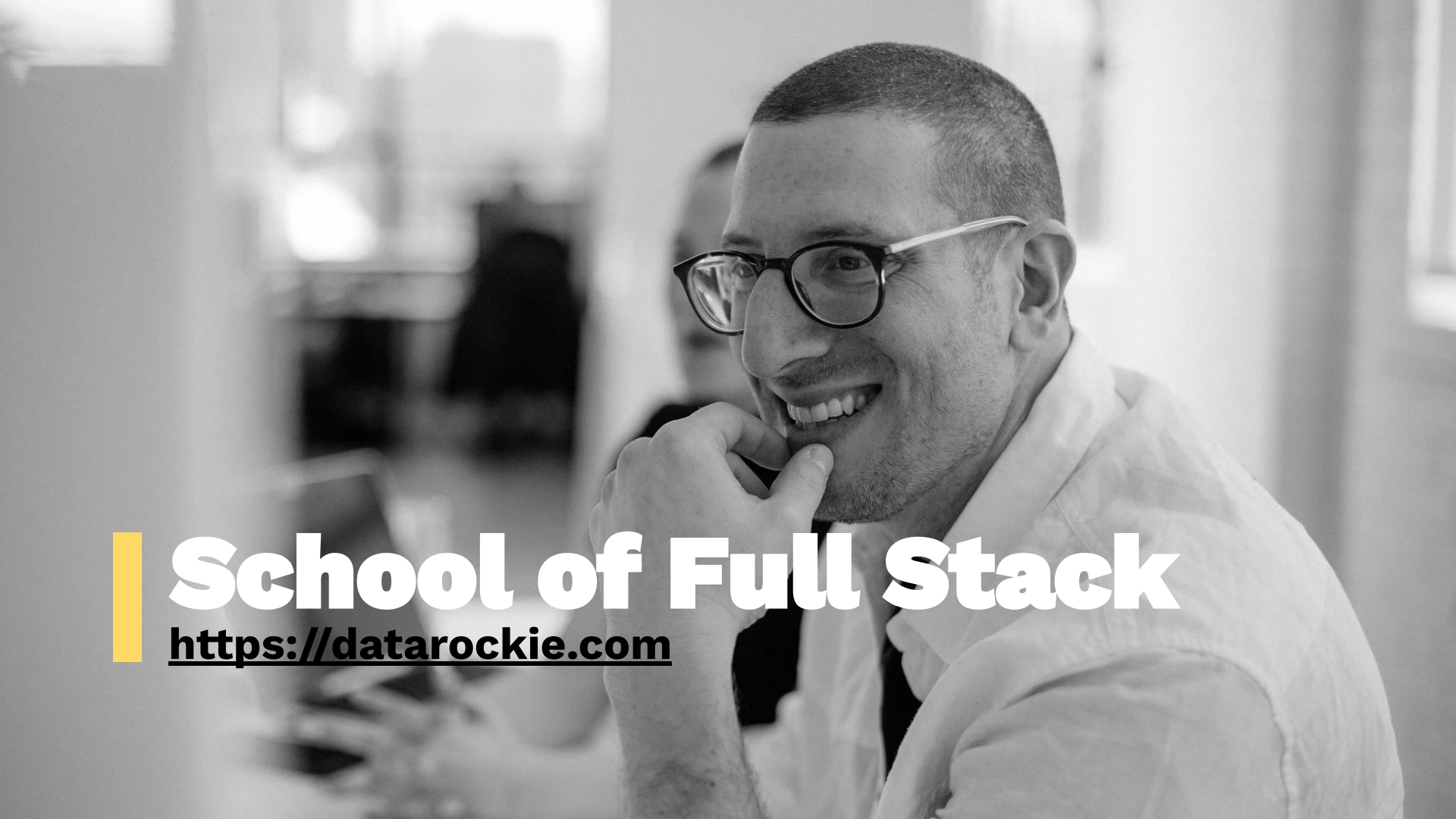
There are seven visualization methods (parameter `method`) in **corrplot** package, named `"circle"`, `"square"`, `"ellipse"`, `"number"`, `"shade"`, `"color"`, `"pie"`.

> Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients.

```
library(corrplot)

## corrplot 0.84 loaded

M <- cor(mtcars)
corrplot(M, method = "circle")
```



School of Full Stack

https://datarockie.com